STATISTICS WORKSHEET-1

MCQ

1. Bernoulli random variables take (only) the values 1 and 0.

a) True b) False

Ans. b) True.

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases? a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned

Ans. b) Central mean Theorem.

3. Which of the following is incorrect with respect to use of Poisson distribution? a) Modeling event/time data b) Modeling bounded count data c) Modeling contingency tables d) All of the mentioned

Ans. b) Modeling bounded count data.

4. Point out the correct statement. a) The exponent of a normally distributed random variables follows what is called the log- normal distribution b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent c) The square of a standard normal random variable follows what is called chi-squared distribution d) All of the mentioned

Ans. c) The square of a standard normal random variable follows what is called chi-squared distribution

5. ______ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

Ans. c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT. a) True b) False

Ans. b) False

7. Which of the following testing is concerned with making decisions using data? a) Probability b) Hypothesis c) Causal d) None of the mentioned

Ans. b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data. a) 0 b) 5 c) 1 d) 10

Ans. a) 0

9. Which of the following statement is incorrect with respect to outliers? a) Outliers can have varying degrees of influence b) Outliers can be the result of spurious or real

processes c) Outliers cannot conform to the regression relationship d) None of the mentioned

Ans. c) Outliers cannot conform to the regression relationship

Subjective answer type questions

10. What do you understand by the term Normal Distribution?

Ans. The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about its mean. In a normal distribution:

The mean (average) and median are equal.

The distribution is bell-shaped and symmetric.

The total area under the curve is equal to 1.

The curve approaches, but never touches, the x-axis as it extends infinitely in both directions.

The shape of the curve is determined by two parameters: the mean (μ) and the standard deviation (σ).

Many natural phenomena and measurements in fields such as statistics, social sciences, natural sciences, and engineering can be approximately described using a normal distribution. The central limit theorem also states that the sampling distribution of the mean of any independent and identically distributed random variables approaches a normal distribution as the sample size increases, regardless of the original distribution of the variables. This makes the normal distribution widely used in statistical inference and hypothesis testing.

11. How do you handle missing data? What imputation techniques do you recommend?

Ans. Handling missing data is a crucial aspect of data analysis. Some common techniques for dealing with missing data include:

Deletion: Removing rows or columns with missing values. This can be appropriate if the missing data is random and does not introduce bias.

Imputation: Filling in missing values with estimated or calculated values. Some common imputation techniques include:

Mean/Median imputation: Replacing missing values with the mean or median of the observed data.

Mode imputation: Replacing missing categorical values with the mode (most frequent value).

Forward fill or backward fill: Propagating the last observed value forward or backward to fill missing values in time series data.

Regression imputation: Predicting missing values using regression models based on other variables.

K-nearest neighbors (KNN) imputation: Filling in missing values based on the values of the nearest neighbors in the feature space.

12. What is A/B testing?

Ans. A/B testing, also known as split testing, is a method used to compare two versions of a product or service to determine which one performs better. It involves dividing users or participants into two groups, where each group is exposed to a different version of the product, advertisement, webpage, or other elements being tested. One group is exposed to the original version while the other group is exposed to a modified version

The purpose of A/B testing is to assess whether the changes made in the treatment version result in a statistically significant difference in outcomes, such as user engagement, conversion rate, click-through rate, or any other key performance metric. By comparing the performance of the two versions, organizations can make data-driven decisions about which version to implement or further iterate upon.

A/B testing is widely used in marketing, product development, web design, and user experience optimization to improve the effectiveness of campaigns, features, and overall user satisfaction.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation of missing data is a commonly used practice, particularly when the missing data is believed to be missing completely at random or missing at random. However, it can introduce bias and underestimate variability, especially if the missingness is non-random or if a large proportion of data is missing. Therefore, while mean imputation is convenient, its acceptability depends on factors such as data characteristics, assumptions, and the impact on analysis results.

while mean imputation is a simple and quick method for handling missing data, its acceptability depends on the specific context of the data and the assumptions being made. It's essential to consider the limitations and potential biases introduced by mean imputation and to explore alternative methods if necessary.

14. What is linear regression in statistics?

Ans. Linear regression is a statistical method used to analyze the relationship between two or more variables. It seeks to find the linear relationship between a dependent variable (the variable being predicted) and one or more independent variables (the predictors). The goal is to create a linear equation that best fits the data, allowing for prediction or understanding of the relationship between the variables.

15. What are the various branches of statistics?

Ans. Statistics is a broad field with several branches that encompass various aspects of data analysis and inference. Some of the main branches of statistics include:

Descriptive Statistics: Involves methods for summarizing and describing characteristics of a dataset, such as measures of central tendency (mean, median, mode) and measures of variability (standard deviation, variance).

Inferential Statistics: Focuses on making predictions or inferences about a population based on a sample of data. This includes hypothesis testing, confidence intervals, and regression analysis.

Probability Theory: Provides the mathematical foundation for statistical methods, including the study of random variables, probability distributions, and stochastic processes.

Biostatistics: Applies statistical methods to biological and health-related data, including clinical trials, epidemiological studies, and medical research.

Econometrics: Applies statistical methods to economic data, often focusing on modeling and analyzing economic relationships and forecasting economic trends.

Actuarial Science: Involves the application of statistical and mathematical methods to assess risk in insurance, finance, and other industries.

Spatial Statistics: Focuses on analyzing spatial data and patterns, including methods for understanding spatial relationships, interpolation, and spatial modeling.

Time Series Analysis: Deals with analyzing data collected over time, such as financial data, weather data, or economic indicators, to identify patterns, trends, and forecast future values.

Multivariate Statistics: Deals with the analysis of datasets with more than two variables, often involving techniques such as principal component analysis, factor analysis, and cluster analysis.

These branches of statistics overlap and intersect with each other, and practitioners often draw from multiple branches depending on the nature of the data and the analysis objectives.